初年次ゼミ「バーチャルリアリティ入門」

音声・音響・音楽のVR

東京大学　工学部　計数工学科
システム情報工学コース第一研究室
猿渡洋・高道慎之介

（2019年4月）

計数システム情報第一研究室

猿渡洋（教授）

専門分野

- 音メディアシステム
- 教師無し最適化
- 統計・機械学習論的信号処理

小山翔一（講師）

専門分野

- 音響信号処理
- 音場再生・伝送（音響ホログラフ）
- スパース信号処理

高道慎之介（助教）

専門分野

- 音声信号処理
- 統計的音声合成
- 声質変換
- 深層学習（DNN）

北村大地（客員研究員）

香川高専

専門分野

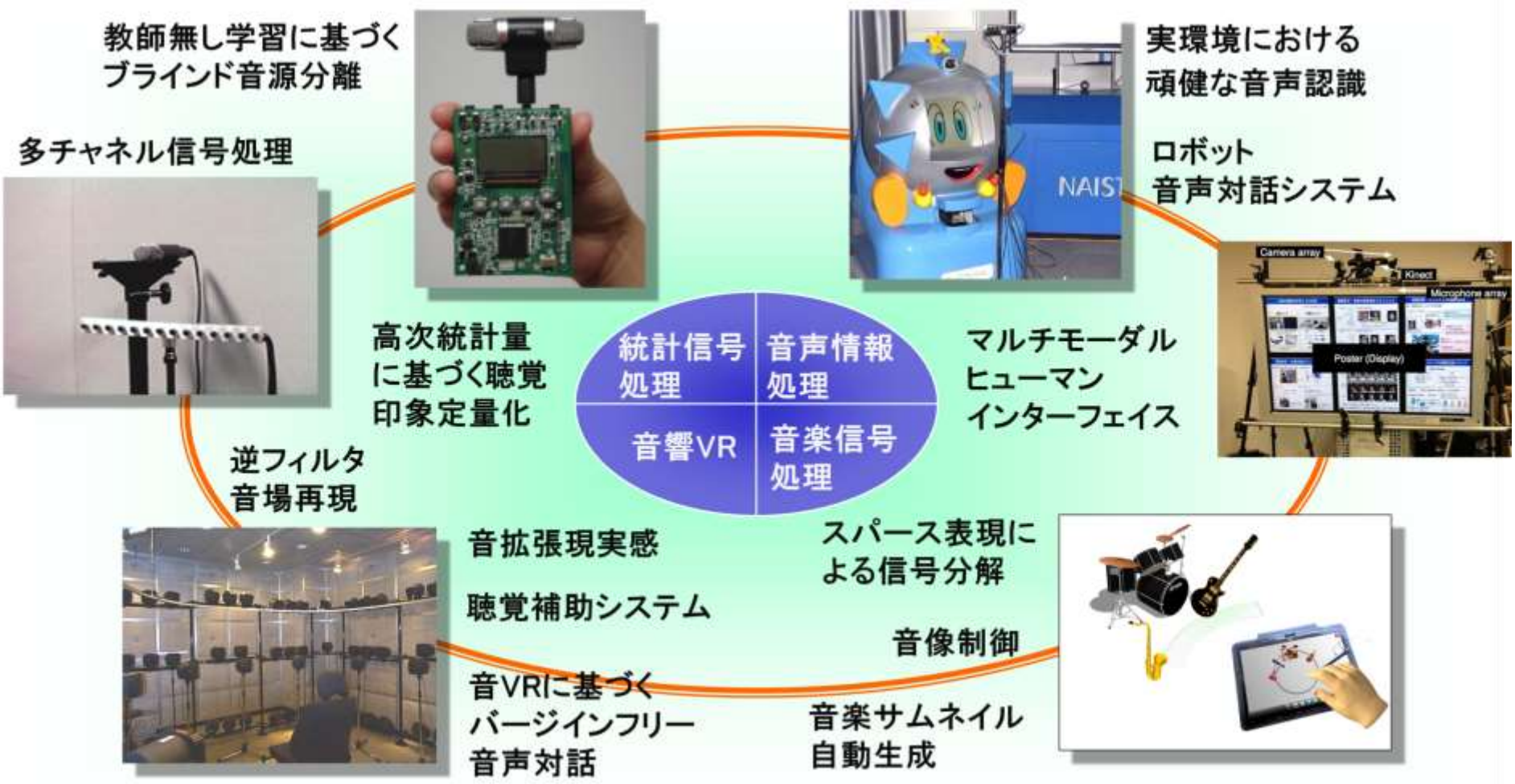
- 音メディア信号処理
- 統計・機械学習論的信号処理
- 音楽信号処理

協力教員　郡山知樹先生
特任研究員　高宗さん
秘書　丹治さん

博士課程学生4名
修士課程学生6＋7名
柏野研学生1名

第一研究室の研究俯瞰図

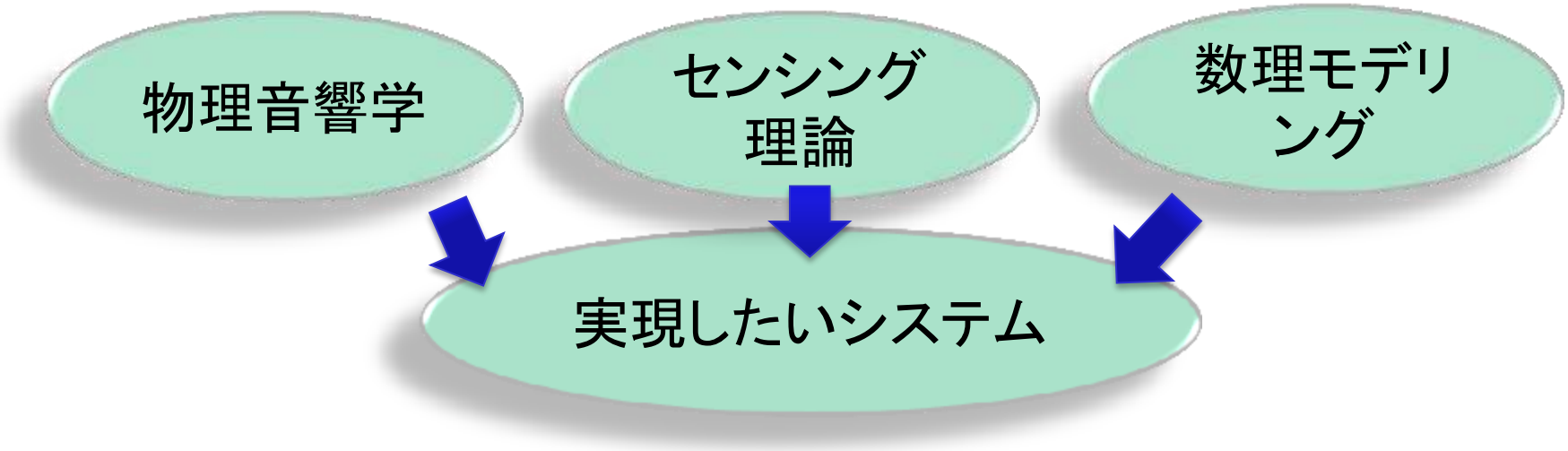
- 音声・音響・音楽メディアに関する信号処理・情報処理
- ヒューマンインターフェイス・コミュニケーションシステムの構築
- 統計的・機械学習論的信号処理、数理最適化問題等を研究



音メディアに関する信号処理研究の魅力

- 音メディアに関する信号処理研究の魅力とは？
  - 自然界の音が持つ無限の多様性（cf. 無線通信信号）
  - 研究のアプローチに多面性あり（決定論的？統計的？）
  - 最後は聴かせてなんぼの評価　⇒　芸術性も併せ持つ

- 「物理世界（波動）と情報世界（抽象）をまたぐ学問」であり、かつそれを「統一的に取り扱うシステム工学」である。
- 対象の多様性ゆえに「なんでもあり」の分野でもある。

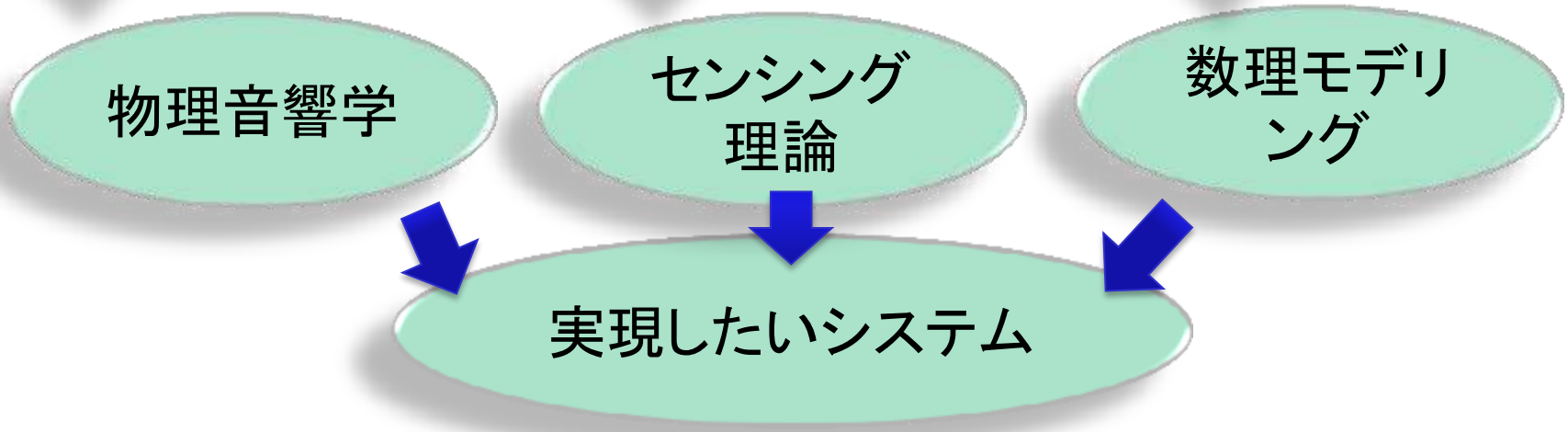


音メディアに関する信号処理研究の魅力

波動方程式
室内音響
伝達関数
音生成過程 etc.

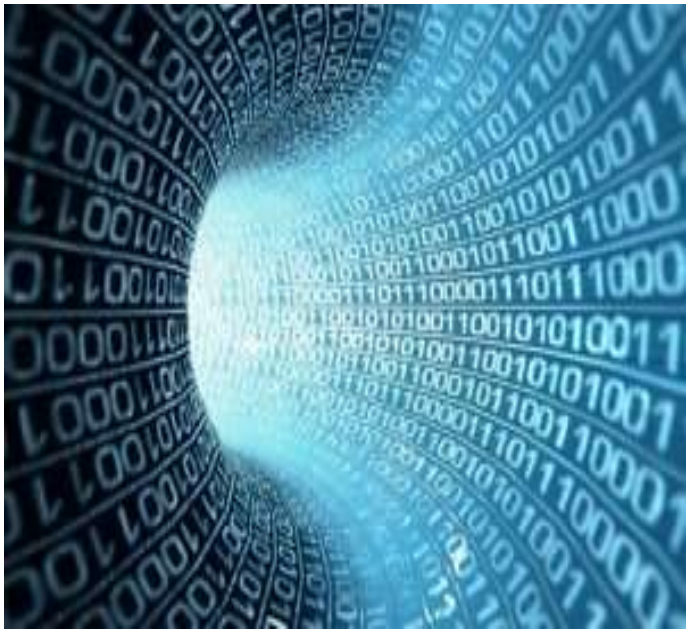
離散サンプリング
フーリエ解析
球面調和解析
圧縮センシング etc.

統計モデリング
最尤・ベイズ推定
機械学習
スパース最適化



ビッグデータ vs. スモールデータ？

時代はビッグデータ！
しかし実際の音波動データは簡単に集められるのか？



むしろ「スモールデータ（一期一会データ）」の研究が必要ではないか？

全てのセンサはネットワークに繋がり「トリリオン（一兆個）センシング」も登場。

しかしばらまかれた音センサは位相情報を失っているので活用できない！

ビッグデータ vs. スモールデータ？

時代はビッグデータ！
しかし実際の音波動データは

全てのセンサはネットワーク
に繋がり「トリリオン（一兆個）

むしろ「スモールデータ（一期一会データ）」の研究が必要ではないか？

しかしばらまかれた音センサは位相情報を失っているので活用できない！

- 1研ではこれら**両方のスケールを意識**した新しい情報処理・機械学習の枠組みを提案します。
- 単に「データ（観測）⇒抽象化（モデリング・認識）」の一方向ではなく、「データ⇒モデリング・認識⇒物理波動の**芸術的**再構築」まで見据えたメディア処理を提案します。

以降では…

- 音VRにおいて必要とされる要素技術の紹介を行う。
- 音メディアは光のような直進性が無く、隠ぺい等を利用できないので、音情景の「分解」・「変換」・「再合成」という3要素が重要になる。

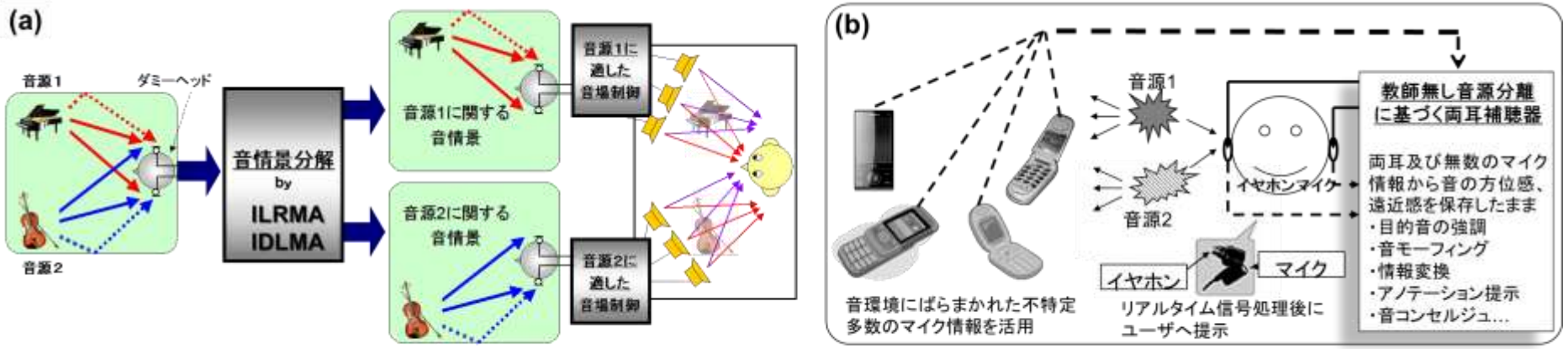


図１．（a）音情景分解・合成システム、（b）音コミュニケーション拡張システム

研究紹介1．ブラインド音源分離

- 音の方向・声質・音量など、事前に何も分かっていなくても、瞬時に音を「聞き分ける」ことの出来るシステムを目指す。

- 独自に開発した高速独立成分分析（ICA）、独立低ランク行列分析（ILRMA）という教師無し数理最適化アルゴリズムに基づいて、音を統計的に独立な成分に分解することにより、別々の音声信号を見つける。

スモールデータ
教師無し最適化
低ランクモデル

音の教師無し分離（AI聖徳太子を造る）

- カクテルパーティー効果
  - 人の聞き分け能力の模擬
  - 補聴器への応用

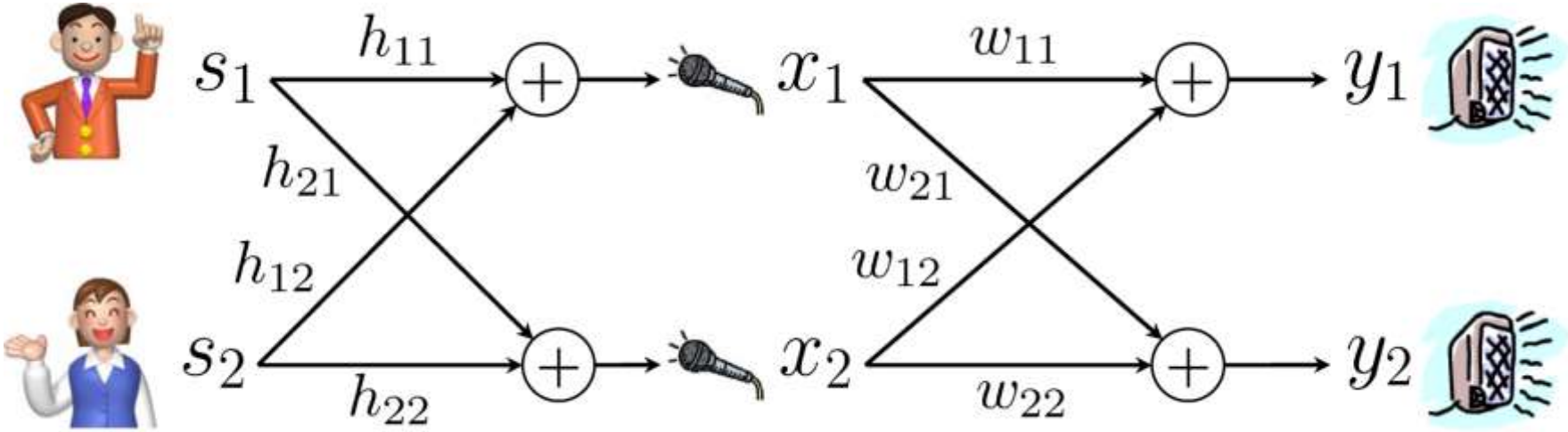
- 音声インターフェイス
  - マイクロホンと口の間の距離が大きくなるにつれて増大してくる妨害音を抑圧･除去
  - 対話IFやロボットへの応用
  - 音監視システム･災害救助ロボット
    （内閣府プロジェクト；2016プレスリリース）

節々にセンサを配置⇒位置不定

- 音楽／楽器音分析
  - ミックスダウン録音の解析
  - 自動採譜、音楽情報処理

ブラインド音源分離（BlindSourceSeparation）

- 混ざり合った信号 x_1, x_2 から元の信号を取り出す
- どのように混ざったかに関する情報 $\boldsymbol{H}$ は利用できない
- 事前トレーニング出来ない⇒ビッグデータではなく**スモールデータ**



実は上記は**2つのことを同時に推定**している

- [空間] 統計的に独立な音源の分類問題（分離行列Wの推定）
- [音源] 各音源が属する確率分布$p(y)$や構造の推定問題

上記を閉形式で解く方法は存在せず凸問題でもない⇒**大変困難！**

ILRMA：音源の独立性と低ランク性に着目したBSS

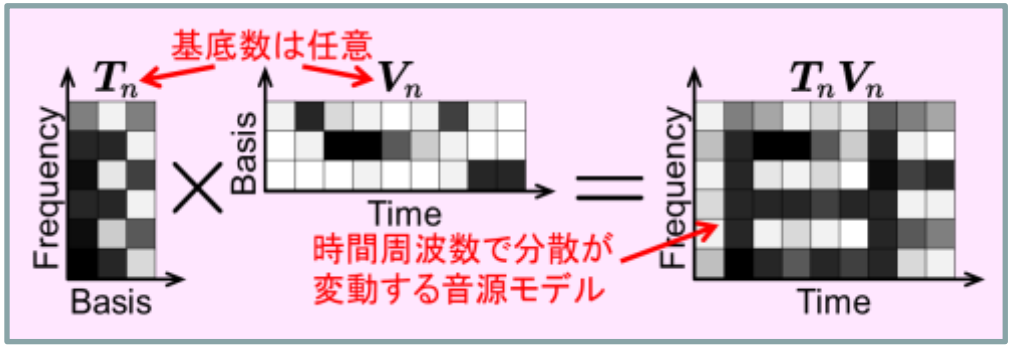
[IEEE Trans. ASLP 2016、IEEE SPJP論文賞・ASJ粟屋賞・JSPS育志賞]

- ILRMAのコスト（対数尤度）関数→これを最小化

$$\mathcal{J} = \sum_{i,j} \left[\sum_m \log \sum_k z_{mk} t_{ik} v_{kj} + \sum_m \frac{|y_{ij,m}|^2}{\sum_k z_{mk} t_{ik} v_{kj}} - 2 \log |\det \boldsymbol{W}_i| \right]$$

音源の低ランク性コスト関数
（音源NMFモデルの推定に寄与）

音源の独立性コスト関数
（空間モデルWの推定に寄与）



$$p(\boldsymbol{y}) = p(y_1)\, p(y_2) \cdots p(y_m)$$

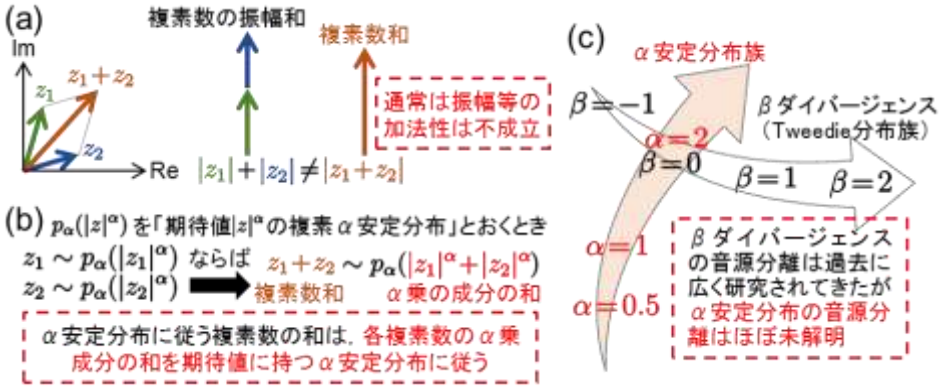
となるWを推定

両者を交互にMajorization-Minimizationアルゴリズムで反復最小化

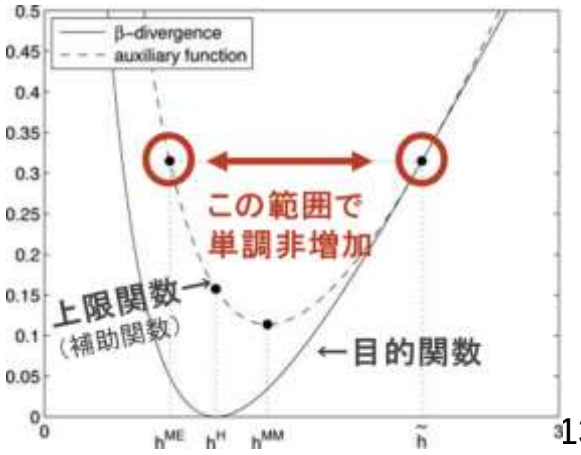
- ✓ コスト値の単調減少性を保証（勾配法には無い特徴）
- ✓ 高速かつ安定な求解法を実現（従来の多入力NMFと比較して2ケタ速い）

様々な研究課題（モデルの多様化・数理解法の開拓）

- 音源生成モデルの多様化
  - 複素波形重ね合わせと整合するα安定分布の導入
    ⇒Student's t-ILRMA [MLSP2017]
  - 複素球状ポアソン分布の導入
    ⇒β=1-divergence最小化ILRMA
    （世界で誰も解けていない！）



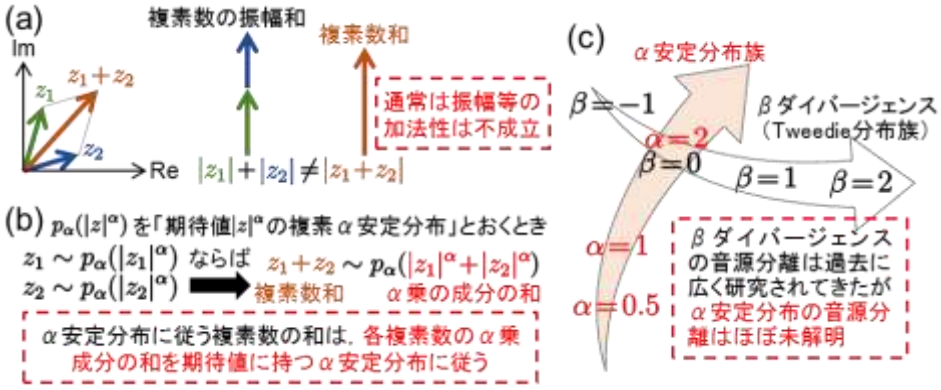
- 座標降下法におけるバリエーション
  - パラメトリックMajorization-Equalizationアルゴリズムによる音源・空間最適化の「バランス」化 [Mitsui et al., CAMSAP2017]



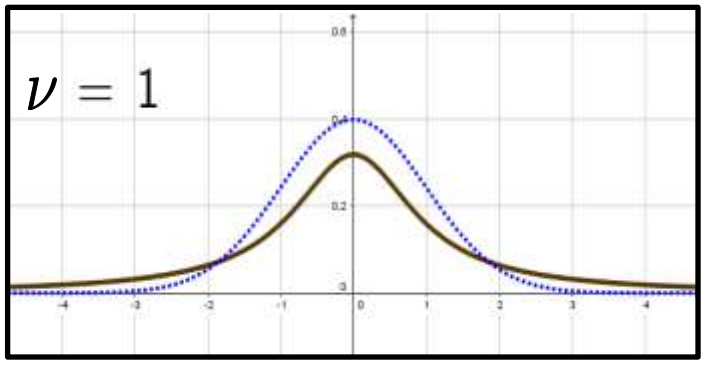
- 識別的な音源NMFモデルの追求
  - 二段階最適化を直接解くことの出来るNMF
  - 教師基底のミスマッチ補正 [遠藤他, 2017]

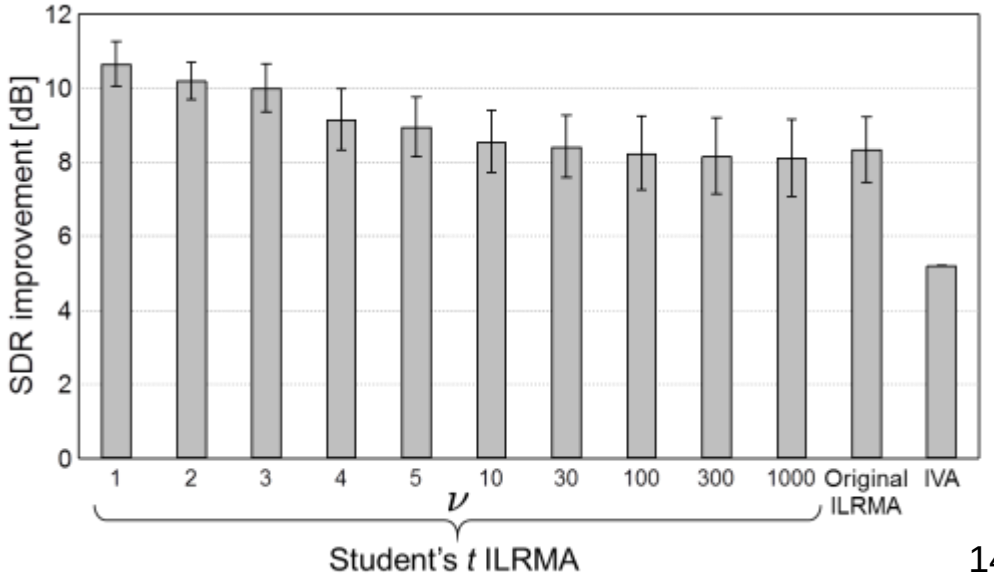
様々な研究課題（モデルの多様化・数理解法の開拓）

- 音源生成モデルの多様化
  - 複素波形重ね合わせと整合するα安定分布の導入
    ⇒Student's t-ILRMA [MLSP2017]
  - 複素球状ポアソン分布の導入
    ⇒β=1-divergence最小化ILRMA
    （世界で誰も解けていない！）



Student's *t*分布

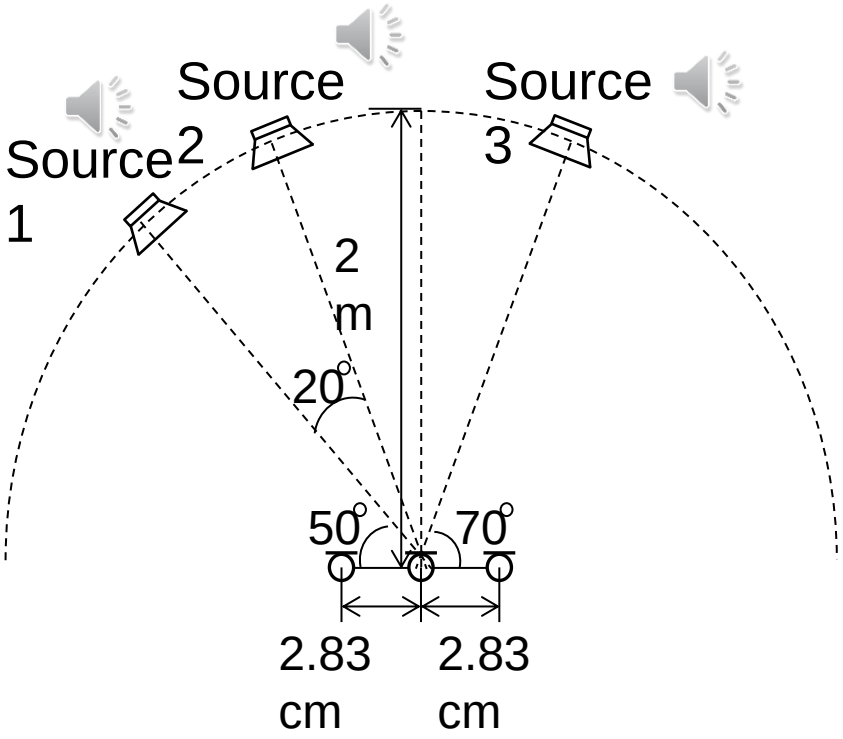




高速ICA、独立低ランク行列分析によるデモ

- リアルタイム音声聞き分け（警察備品に採用）

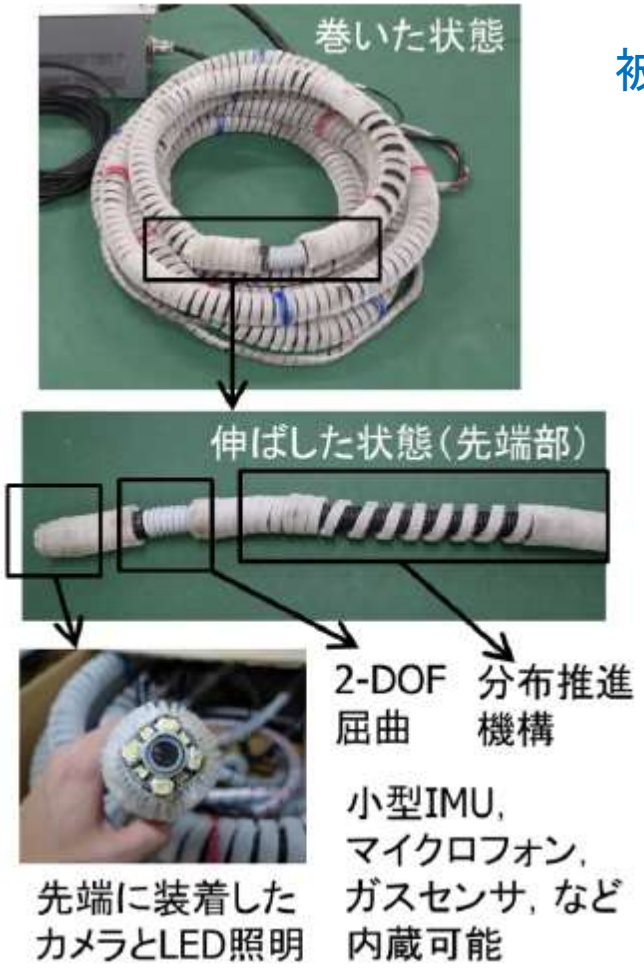
- ドラム、弦楽器、音声からなる複合音の分離

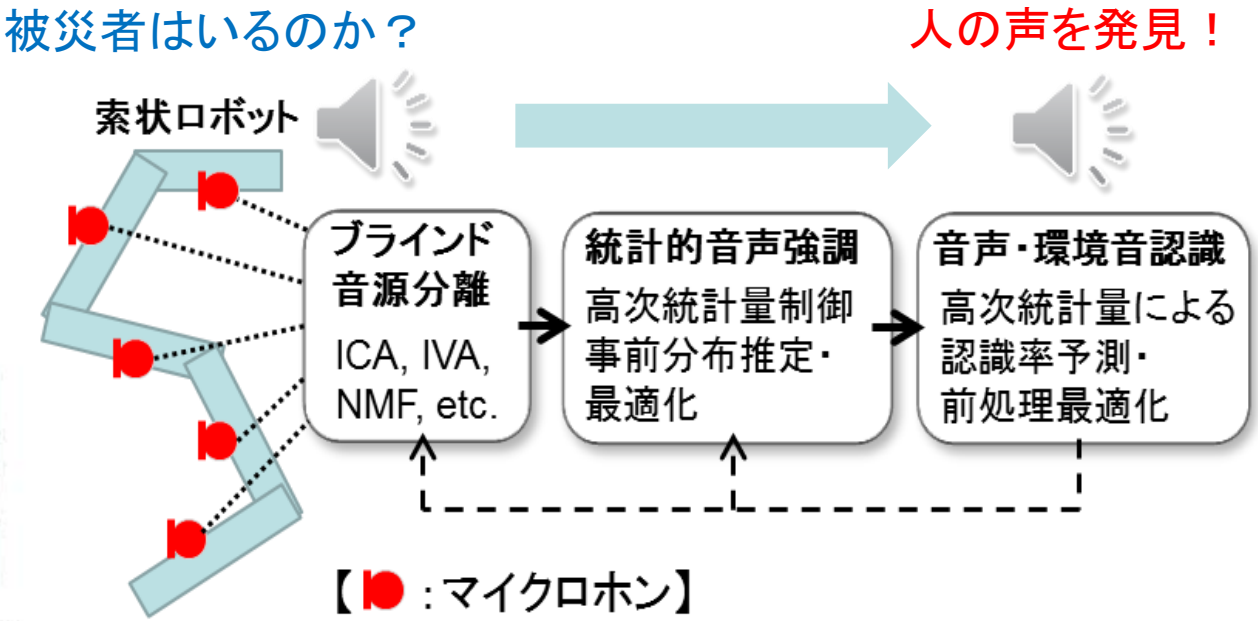


内閣府ImPACT災害対応タフロボット [2016年6月プレスリリース]

■災害時の倒壊家屋に入り込んで被災者発見

■環境音認識による状況把握・救助支援





いかなる曲がりくねった形状においてもマイク同士が協調して騒音の中から被災者の声を見つけ出す

独立深層学習行列分析

Independent Deeply Learned Matrix Analysis
(IDLMA: 発音はアイドルエムエー)

ILRMAにおける問題点：音源の低ランク性？

$$\mathcal{J}_{\mathrm{ILRMA}} = \frac{1}{J}\sum_{i,j,n}\left[\log\sum_{l} t_{il,n}v_{lj,n} + \frac{|\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2}{\sum_{l} t_{il,n}v_{lj,n}}\right] - \sum_{i}\log|\det \boldsymbol{W}_i|^2$$

音源モデル（低ランク性）

音源によっては低ランク性が
成り立たない場合がある

ならば！

事前に学習データを用いて音源モデル
の分散を推定する写像を作る

空間モデル（音源間が独立）

音源・マイク位置，部屋の形状，
残響時間などの膨大な物理要因に依存

学習データの用意は非現実的
ブラインドに推定

ILRMAにおける問題点：音源の低ランク性？

$$\mathcal{J}_{\mathrm{ILRMA}} = \frac{1}{J}\sum_{i,j,n}\left[\log\sum_{l} t_{il,n}v_{lj,n} + \frac{|\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2}{\sum_{l} t_{il,n}v_{lj,n}}\right] - \sum_{i}\log|\det \boldsymbol{W}_i|^2$$

音源モデル（低ランク性）

空間モデル（音源間が独立）

- 深層学習（DNN）による強力なモデリング能力を活用する！
- 今まで培ってきた「教師あり音源分離（例：教師ありNMF）」の技術を昇華させる形で研究を発展できる。
- 急速に発展するDNN研究を我々ならではの視点で拡張する。

事前に学習データを用いて音源モデルの分散を推定する写像を作る

学習データの用意は非現実的

ブラインドに推定

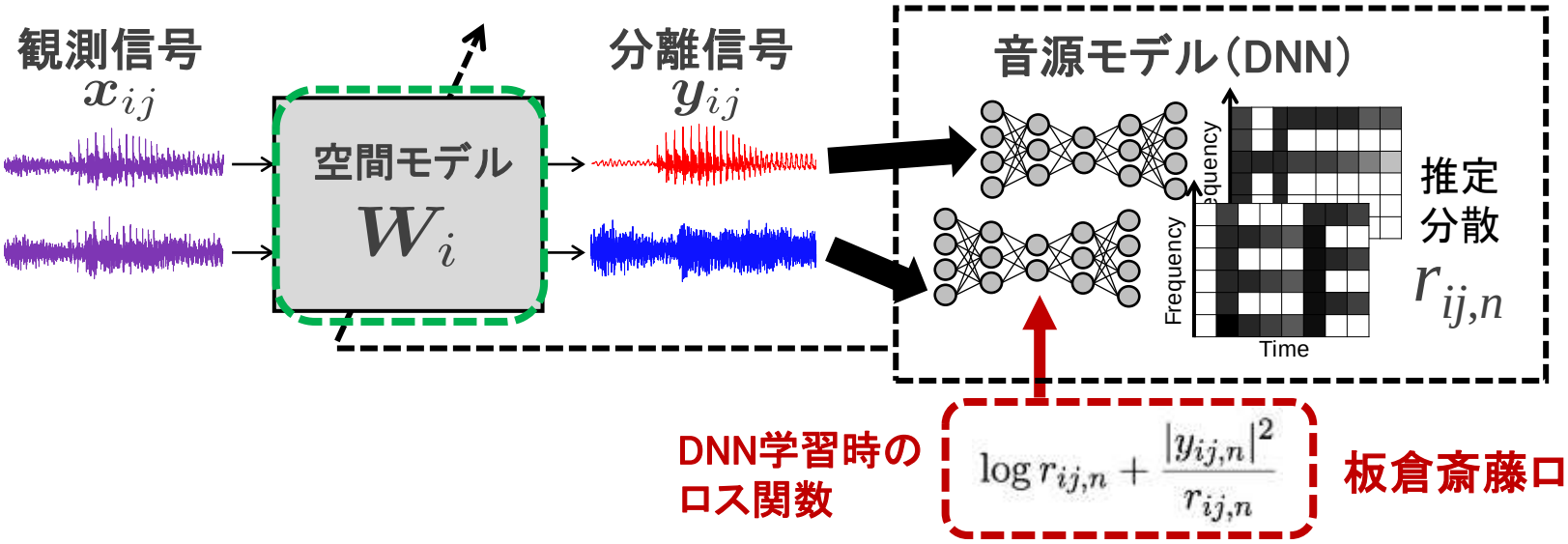
提案手法：DNN音源モデルによる最尤推定

■ 独立深層学習行列分析（IDLMA）

$$y_{ij,n} = \boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}$$

$$\mathcal{J} = \frac{1}{J} \sum_{i,j,n} \left(\log r_{ij,n} + \frac{|\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}|^2}{r_{ij,n}} \right) - \sum_{i} \log |\det \boldsymbol{W}_i|^2$$

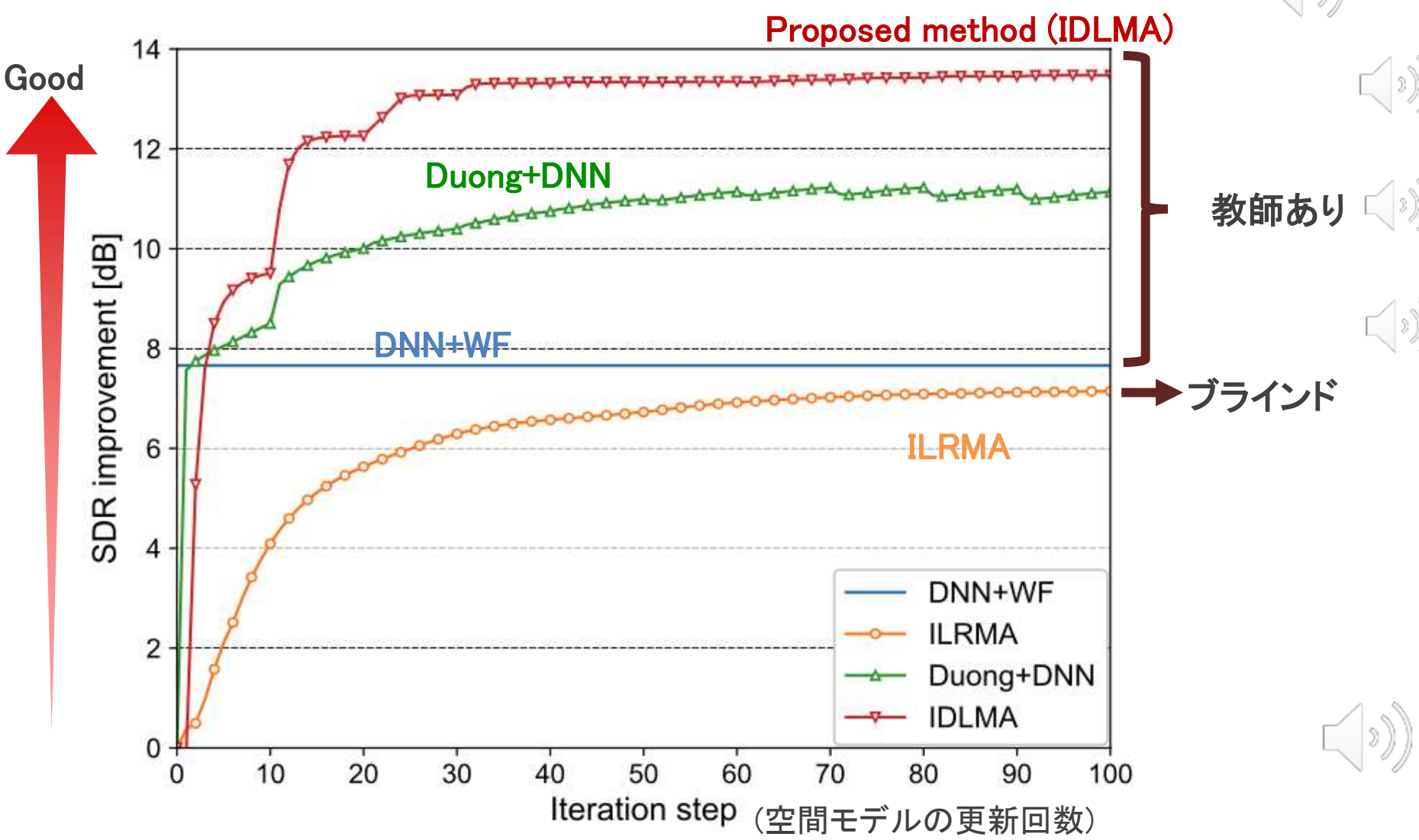
音源モデル（DNN）　⟷　交互に最適化　⟷　空間モデル（音源間が独立）



■ **空間モデル**：各音源が統計的に**独立**となる分離行列を推定

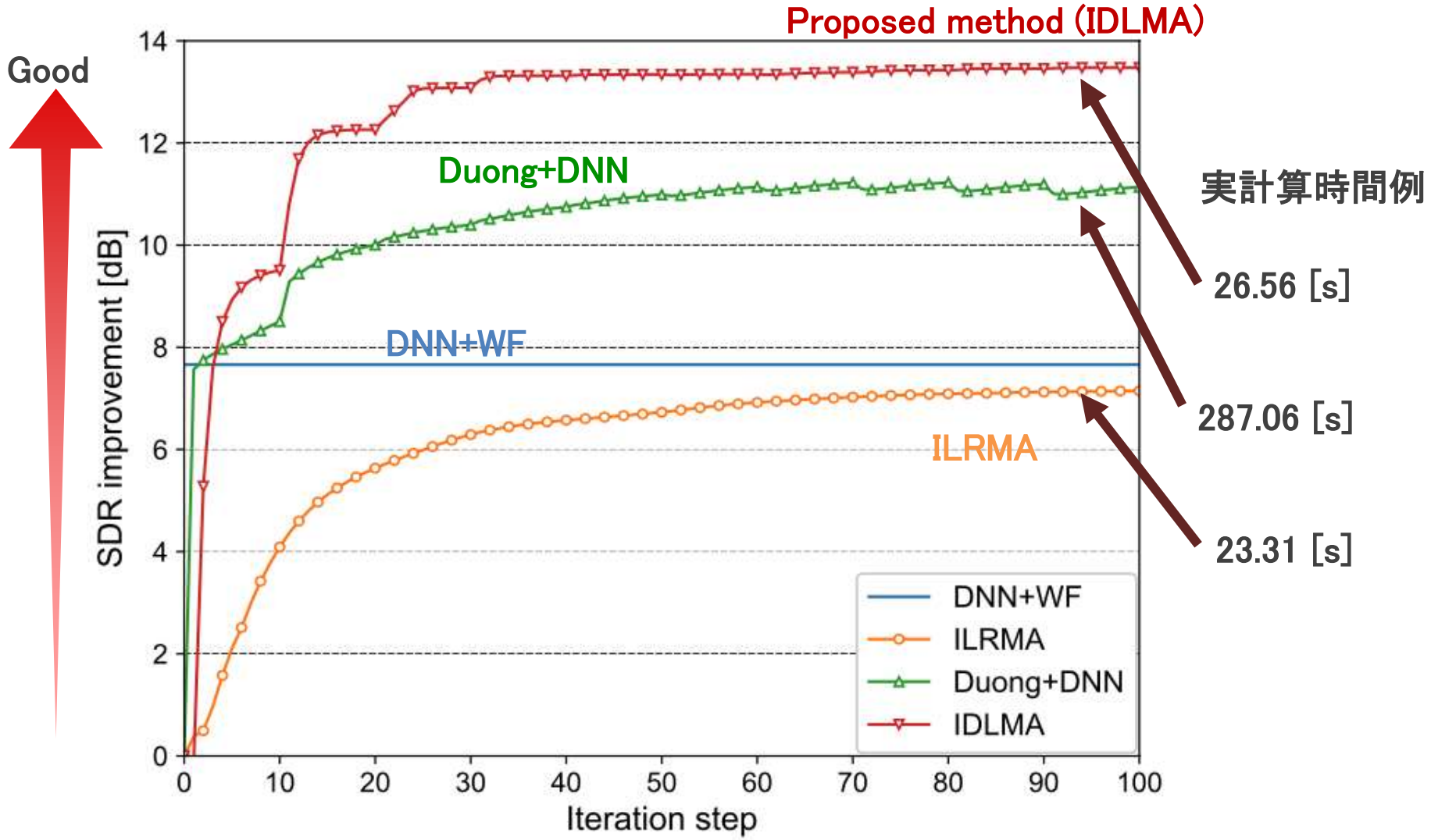
■ **音源モデル**：J を最小化するような分散 $r_{ij,n}$ を推定するDNNを各音源ごとに構成

簡易実験結果（2018年3月の音響学会で発表）



■ 10回に1回 DNNで分散行列を更新

簡易実験結果（2018年3月の音響学会で発表）



■ 10回に1回 DNNで分散行列を更新

研究紹介2. 音楽信号解析

- 様々な楽器がまじりあった音楽信号の中から、自分の好きな楽器を見つけ出し、自分の好みのリミックス版を製作する。

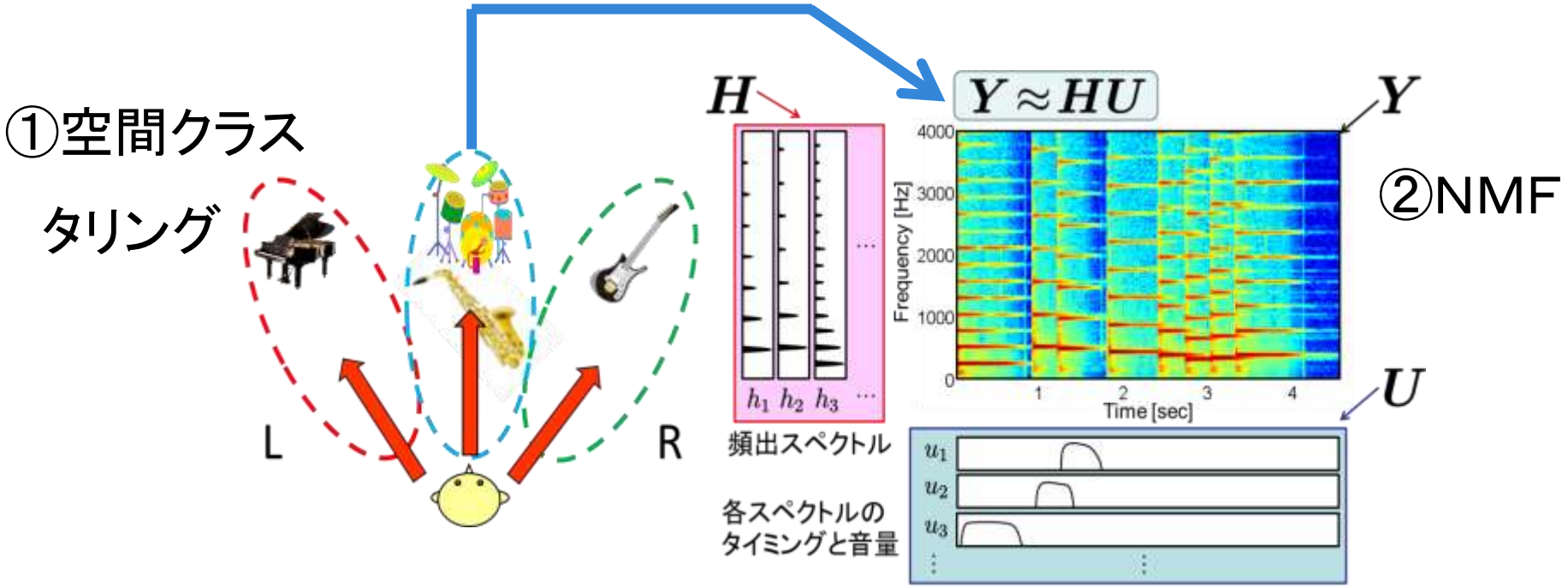
- 非負値行列因子分解（NMF）という数理アルゴリズムに基づいて、音を事前に学習した「より簡単（低ランクかつ疎：スパース）な頻出パターン」に分解することにより、信号を解析する。

スパース分解
低ランクモデル
半教師有り学習

スパース・アンチスパース信号表現に基づく多チャネル音楽信号分離

[IEEE Trans. ASLP 2015, SIP若手奨励賞、日本音響学会学生優秀発表賞、TAF論文賞]

- 多チャネル音楽信号を効率よく分解するため、空間クラスタリングとスペクトル頻出パタン分解（非負値行列因子分解：NMF）を組み合わせた手法を提案している。
- 本手法では、空間クラスタリングでの分類エラーを学習基底で補修する機能が備わっているため、分離に必要なスパース性と補修に必要なアンチスパース性との間でトレードオフが生じる。
- そこで、各音源に合ったスパース性を選択する機能を導入した。



多チャネル音楽信号分離デモ

- 実際の演奏曲を教師有りNMFで分解してみた。

原曲

教師１

分離音１

教師２

分離音２

多チャネル音楽信号分離デモ

- プロレコーディングに対応できる品質を目指して。

原曲（プロ演奏）

Saxのみを抜いた伴奏部分

YAMAHA

サックス奏者が消えた!?

研究紹介3．音質の定量化

- 各種の統計的音声推定を行う場合、非常に不愉快な人工雑音が残留し、出力音の「聴感的な印象」を下げてしまう。

- 統計推定手法ごとに「聴感的印象」は異なる。つまり、統計的推定には「音の個性」がある。我々は芸術的観点から統計的推定問題を眺める。

- 聴感的印象を数値化し、その値が不動となるパラメータの存在を世界で初めて発見した。

非線形システム
高次統計量解析
聴覚印象定量化

非線形処理の問題点：アーチファクトの発生

◆一般に、非線形雑音抑圧信号処理において、不快なアーチファクト（これはミュージカルノイズと呼ばれる）が発生する。これは処理を「人が聴く」用途へ適用する際に、大きな問題となってしまう。

◆このアーチファクトは、各種統計推定方式によって異なる。例えば、振幅スペクトル最尤推定、複素スペクトル最小二乗推定、振幅スペクトルベイズ推定の順に軽減される。しかし、本質的な改善には至っていない。

◆本来、ミュージカルノイズに関しては、数理解析がほとんどなされていないという現状がある。よって、

(1) まずミュージカルノイズの定量指標を定める必要がある。

(2) 次に、ミュージカルノイズを低減する信号処理を開発する。

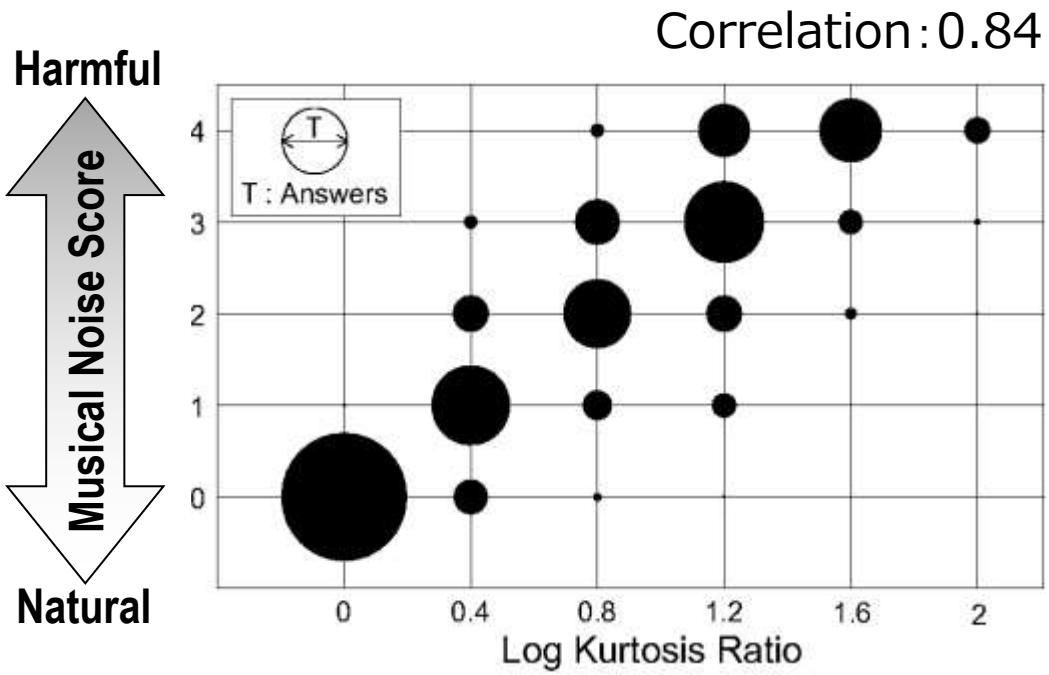
カートシス比：聴感印象は4次統計量で測る

■ 処理前後のカートシス変化比（源信号尖度で正規化）

$$\text{kurtosis ratio} = \frac{\text{kurt}_{\text{proc}}}{\text{kurt}_{\text{org}}}$$

kurt_{org}　入力信号の尖度
$\text{kurt}_{\text{proc}}$　出力信号の尖度

カートシス比が1　⇒　ミュージカルノイズ無し
カートシス比が1より大　⇒　ミュージカルノイズ多い



各種信号推定における誤差の高次モーメント

振幅スペクトル最尤推定 [EURASIP JASP 2010]

$$\mu_m^{\mathrm{SS}} = \theta^m \mathcal{M}(\alpha, \beta, \eta, m)$$

$$\mathcal{M}(\alpha, \beta, \eta, m) = \frac{1}{\Gamma(\alpha)} \sum_{l=0}^{m} (-\beta\alpha)^l \frac{\Gamma(m+1)}{\Gamma(l+1)\Gamma(m-l+1)} \Gamma(\alpha+m-l, \beta\alpha) + \eta^{2m} \frac{\gamma(\alpha+m, \beta\alpha)}{\Gamma(\alpha)}$$

$\gamma(\alpha, z) = \int_0^z t^{\alpha-1} \exp(-t) dt$ ：第1種不完全ガンマ関数

$\Gamma(\alpha, z) = \int_z^\infty t^{\alpha-1} \exp(-t) dt$ ：第2種不完全ガンマ関数

一般化ウィーナフィルタ [IEEE Trans. ASLP 2011]

$$\mathcal{M}_{\mathrm{QPWF}}(\alpha, \beta, m, \xi, \eta) = \frac{1}{\Gamma(\alpha)} \int_0^\infty \frac{t^{(\xi\eta+1)m+\alpha-1}}{\left\{ t^{\frac{\xi}{2}} + \beta \frac{\Gamma(\alpha+\frac{\xi}{2})}{\Gamma(\alpha)} \right\}^{2m\eta}} \exp(-t) dt$$

各種信号推定における誤差の高次モーメント

ベイジアンMMSE音声振幅推定 [IWAENC2012]

$$\begin{cases} \mu_1 = & K_1(|Y(f,\tau)|^2) \\ \mu_2 = & K_2(|Y(f,\tau)|^2)+K_1^2(|Y(f,\tau)|^2) \\ \mu_4 = & K_4(|Y(f,\tau)|^2)+4K_3(|Y(f,\tau)|^2)K_1(|Y(f,\tau)|^2) \\ & +3K_2^2(|Y(f,\tau)|^2)+6K_2(|Y(f,\tau)|^2)K_1^2(|Y(f,\tau)|^2) \\ & +K_1^4(|Y(f,\tau)|^2) \end{cases}$$

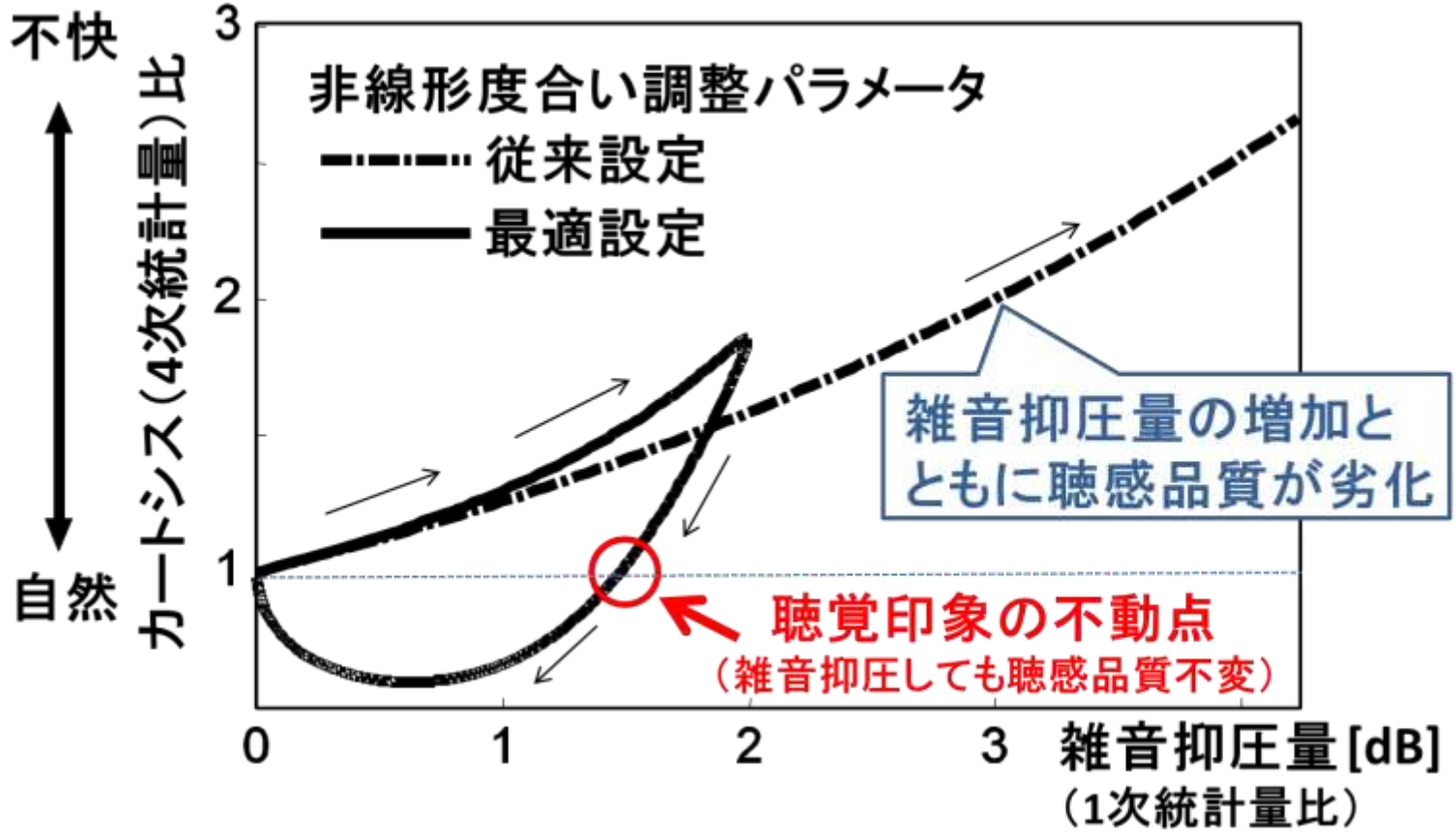
where $K_m(|Y(f,\tau)|^2) =$

$$\sum_{i=1}^{2}(-1)^{i-1}(1-\alpha)^{im}(R(\beta)\eta\theta)^m Z_i(m)K_m((\xi^{\mathrm{ml}}(f,\tau))^i),$$

$$\begin{cases} K_1((\xi^{\mathrm{ml}}(f,\tau))^i) = & \mu_{i[SS]} \\ K_2((\xi^{\mathrm{ml}}(f,\tau))^i) = & \mu_{2i[SS]}-\mu_{i[SS]}^2 \\ K_3((\xi^{\mathrm{ml}}(f,\tau))^i) = & \mu_{3i[SS]}-3\mu_{2i[SS]}\mu_{i[SS]}+2\mu_{i[SS]}^3 \\ K_4((\xi^{\mathrm{ml}}(f,\tau))^i) = & \mu_{4i[SS]}-4\mu_{3i[SS]}\mu_{i[SS]}-3\mu_{2i[SS]}^2 \\ & +12\mu_{2i[SS]}\mu_{i[SS]}^2-6\mu_{i[SS]}^4 \end{cases}$$

$$\mu_{m[SS]} = \sum_{l=0}^{m}(-\eta)^l \frac{\Gamma(m+1)\Gamma(\eta+m-l,\eta)}{\Gamma(\eta)\Gamma(l+1)\Gamma(m-l+1)}.$$

高次統計量空間での不動点の発見



ミュージカルノイズフリー信号処理 [IEEE Trans. ASLP 2012、ASJ板倉賞・市村賞]

- カートシス比不動点の存在は聴覚印象の不動点を表す
- 一次統計量の増分が少ない場合は本処理を繰り返せば良い

統計推定における音色の差を聞いてみよう！

白色ノイズの場合

観測音 

最尤推定 

ベイズ推定 

ミュージカル
ノイズフリー 

人ごみノイズの場合

観測音 

最尤推定 

ベイズ推定 

ミュージカル
ノイズフリー 

どの推定方式が「自然」だと感じましたか？

研究紹介4．音場再現

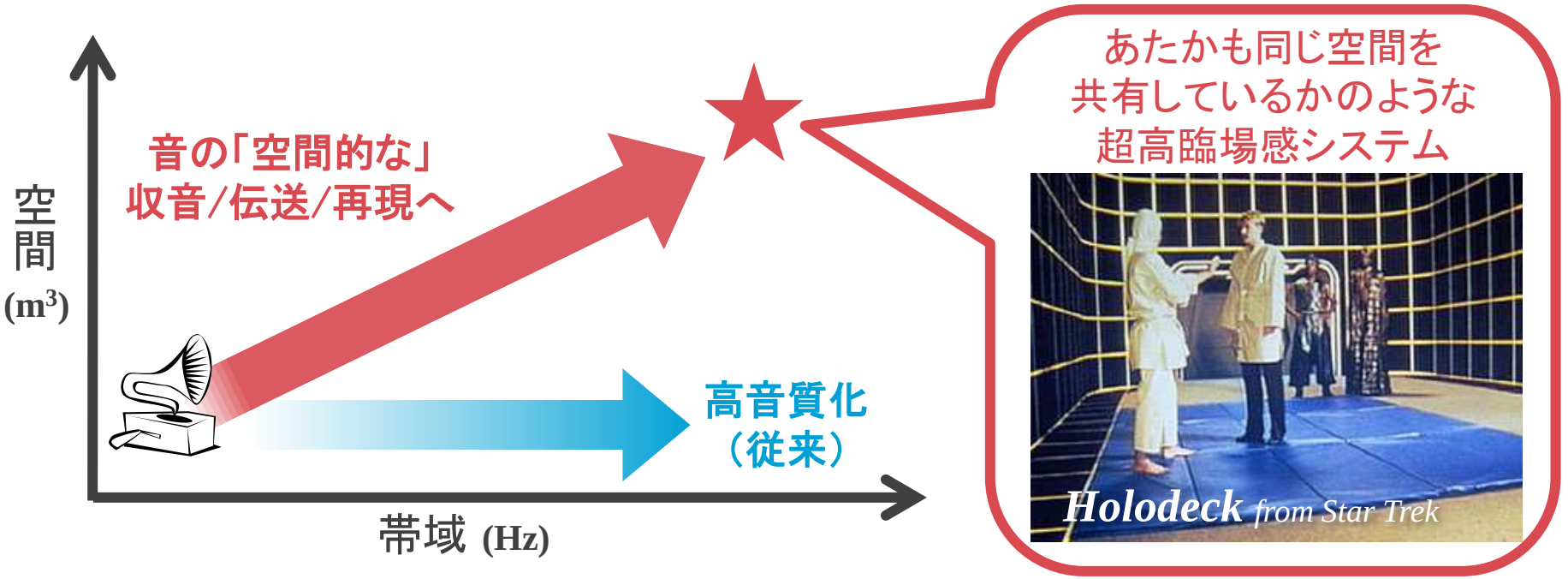
- ステレオや5.1chサラウンドなどの従来の音の再生技術と異なり、広い領域で音の空間そのものを再現する。

- 音のホログラフを実装し、音バーチャルリアリティや音拡張現実感システムを実現する。

- 音場を「スパース（疎）」な性質に基づいて分解する。

球面調和解析
スパース最適化
音VR・AR

超高臨場感音響再生技術に向けて



テレプレゼンス

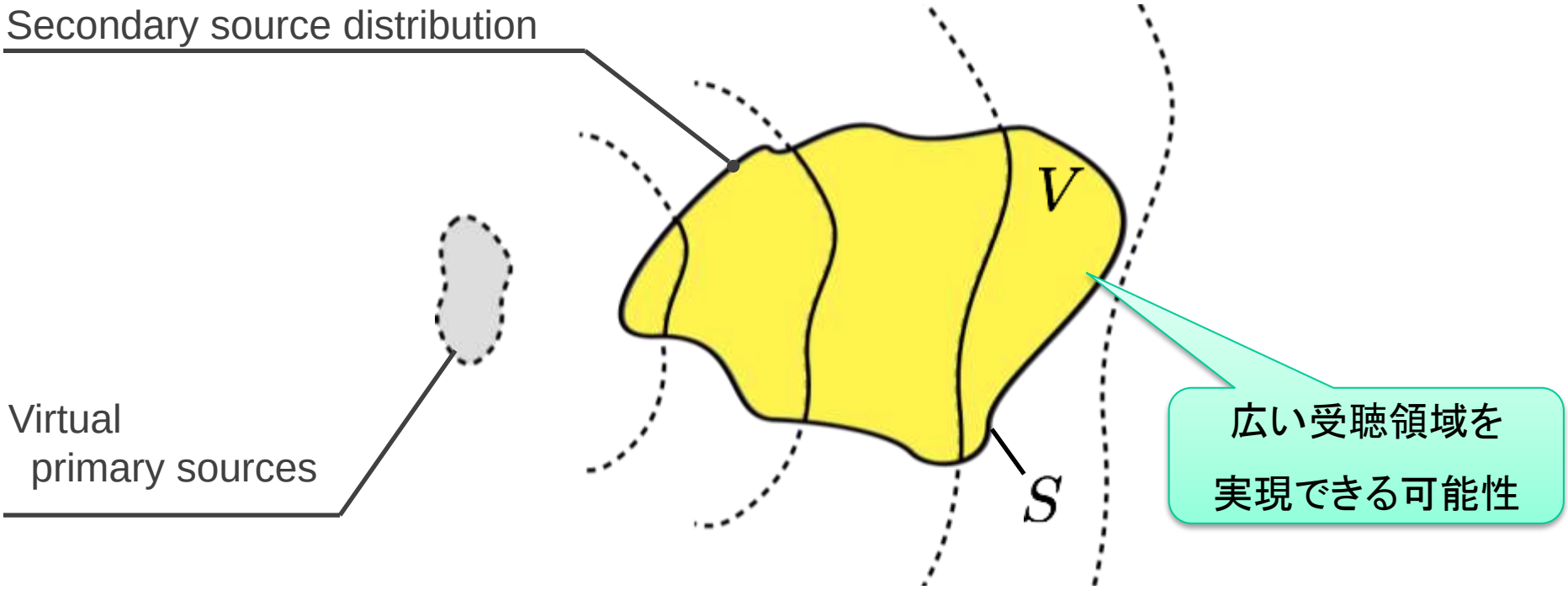
コンテンツ配信

家庭

劇場

音場再現による高臨場音響再生

音場そのものを物理的に再現



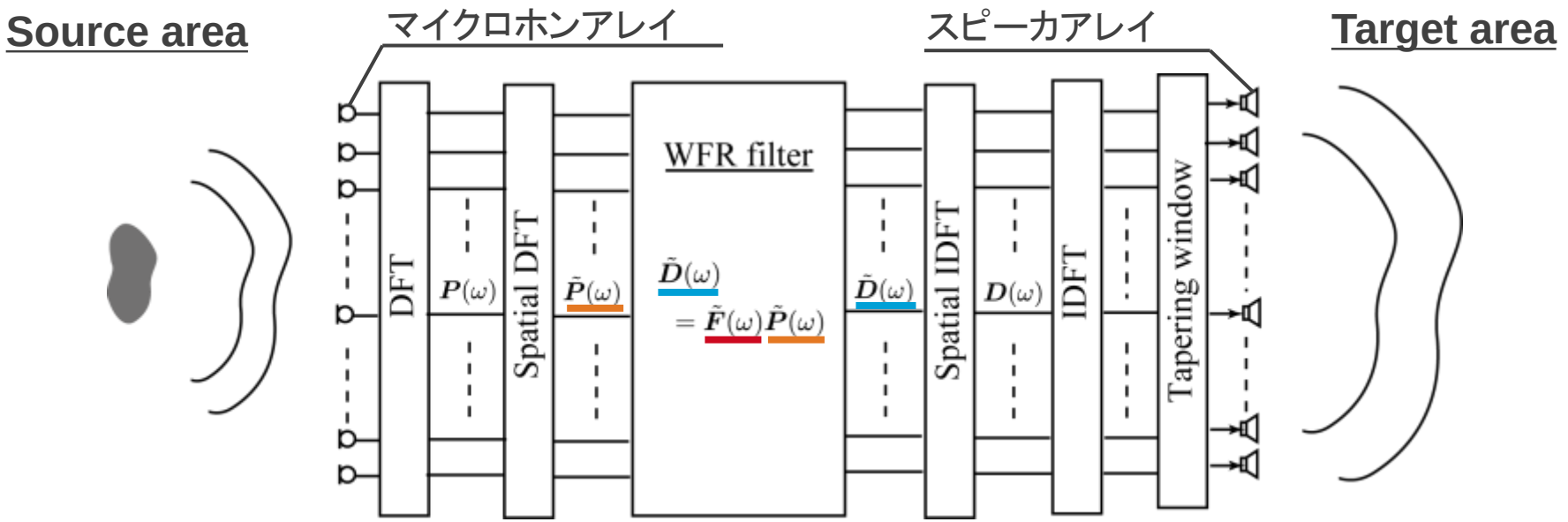
- 対象領域V内の音場を，境界面S上に配置した二次音源（＝スピーカ）を用いて，所望の音場と一致させる

➡ **音場再現問題**

直線状アレイのためのWFRフィルタ

[音響学会板倉賞]
[テレコムシステム技術賞]

実装上は時空間の2次元FIRフィルタ畳み込み：**波面再構成（WFR）フィルタ**



$$\tilde{D}(\omega) = \tilde{F}(\omega)\tilde{P}(\omega)$$

スピーカ駆動信号の
時空間スペクトル

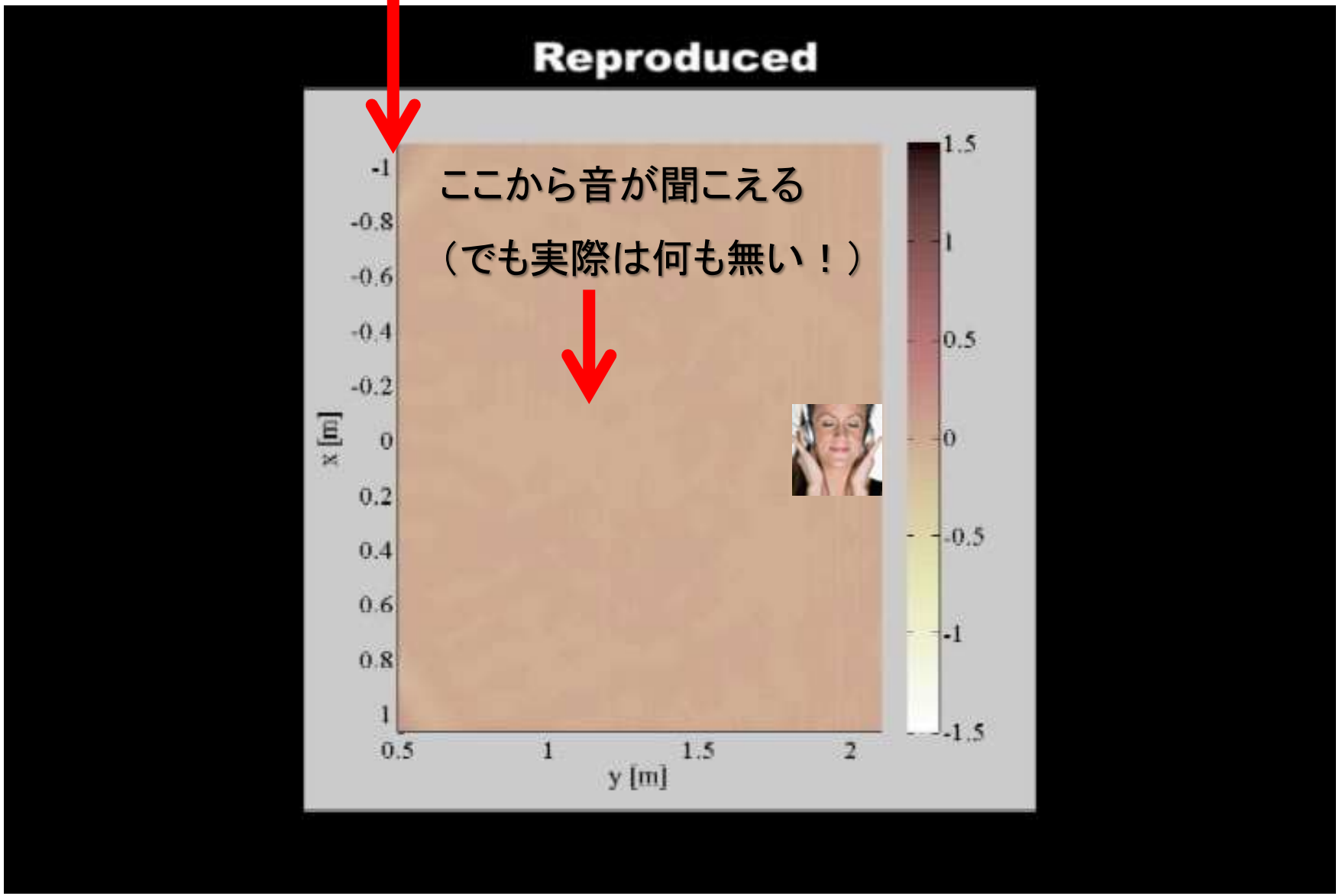
マイクロホン信号の
時空間スペクトル

WFRフィルタ：各要素が以下で決まる対角行列

$$\tilde{F}_i(\omega) = -4j \frac{e^{j\sqrt{k^2 - k_{x,i}^2}\, y_{\mathrm{ref}}}}{H_0^{(1)}\left(\sqrt{k^2 - k_{x,i}^2}\, y_{\mathrm{ref}}\right)}$$

仮想再現音場例1（音源がスピーカの手前）

物理的なスピーカ列はここ



スパース表現に基づく音場分解

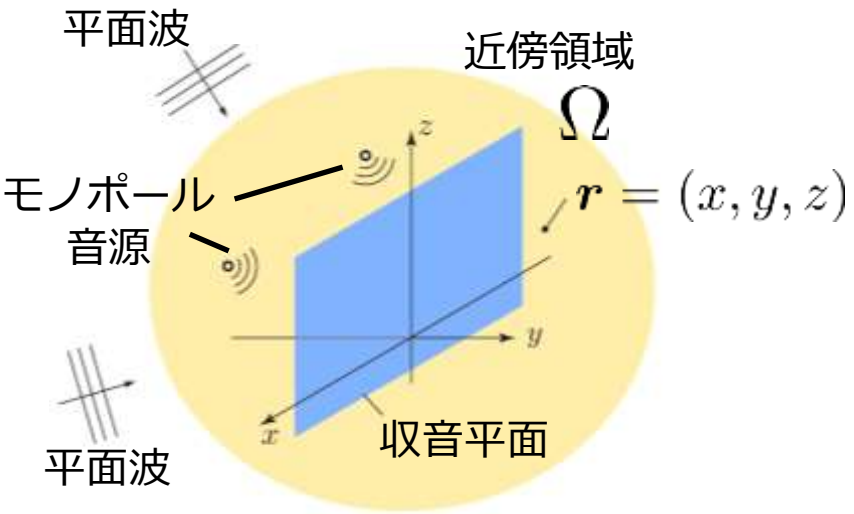
音場の生成モデル：モノポール音源と平面波の重ねあわせでモデル化

[Koyama, 2014]

非斉次・斉次
Helmholtz 方程式による表現

$$(\nabla^2 + k^2)p(\boldsymbol{r},\omega) = \begin{cases} -Q(\boldsymbol{r},\omega), & \boldsymbol{r} \in \Omega \\ 0, & \boldsymbol{r} \notin \Omega \end{cases}$$

$Q(\boldsymbol{r},\omega)$: モノポール音源の分布



このHelmholtz方程式の解は，**非斉次項**と**斉次項**の和として書ける．

$$\begin{aligned} p(\boldsymbol{r},\omega) &= p_i(\boldsymbol{r},\omega) + p_h(\boldsymbol{r},\omega) \\ &= \int_{\boldsymbol{r}' \in \Omega} Q(\boldsymbol{r}',\omega) G(\boldsymbol{r}|\boldsymbol{r}',\omega) d\boldsymbol{r}' + p_h(\boldsymbol{r},\omega) \end{aligned}$$

Green関数

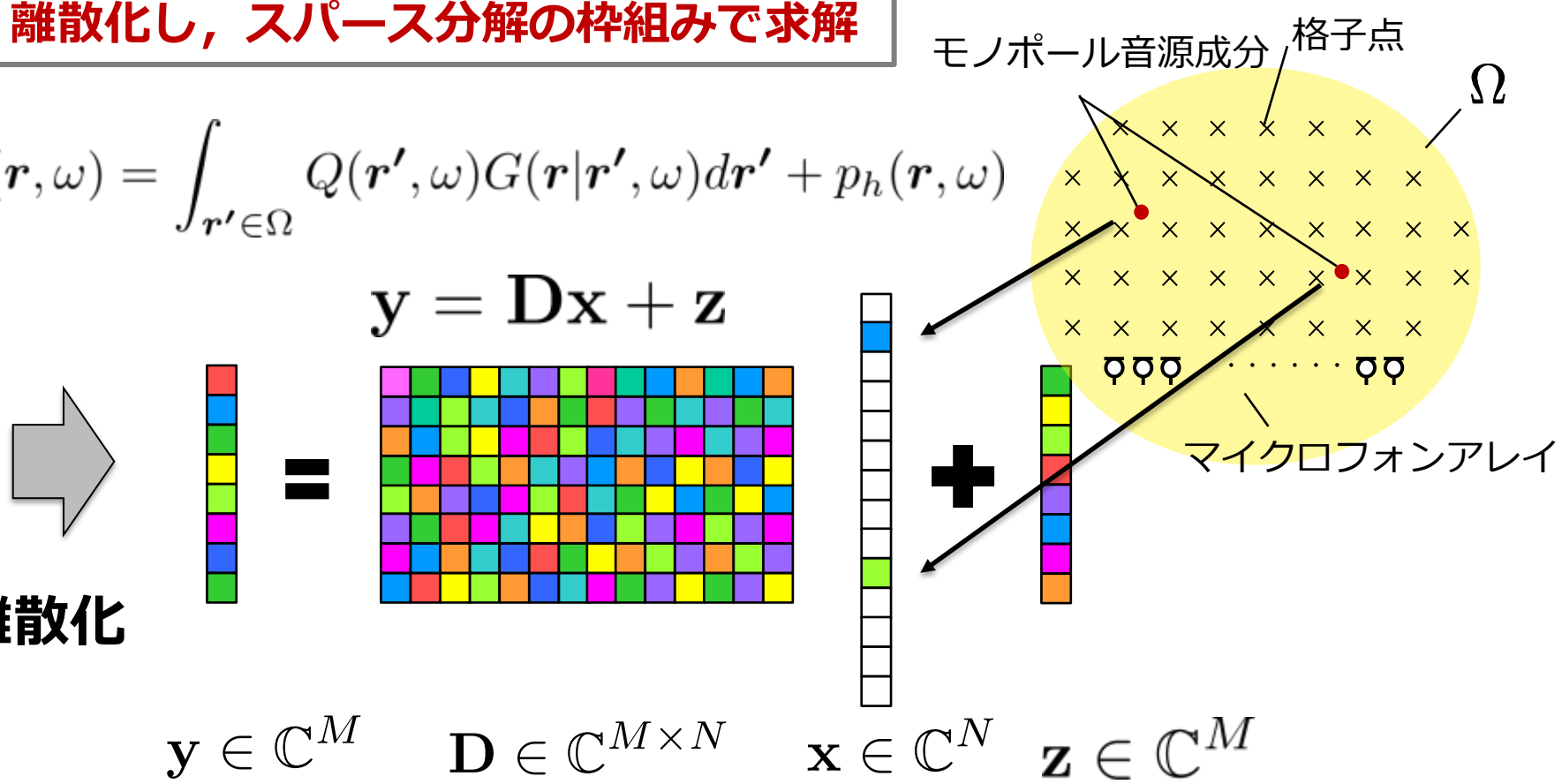
モノポール音源由来の**非斉次項**

平面波由来の**斉次項**

スパース表現に基づく音場分解

離散化し，スパース分解の枠組みで求解

$$p(\boldsymbol{r},\omega) = \int_{\boldsymbol{r}'\in\Omega} Q(\boldsymbol{r}',\omega)G(\boldsymbol{r}|\boldsymbol{r}',\omega)d\boldsymbol{r}' + p_h(\boldsymbol{r},\omega)$$



$\mathbf{y} \in \mathbb{C}^M$ $\mathbf{D} \in \mathbb{C}^{M\times N}$ $\mathbf{x} \in \mathbb{C}^N$ $\mathbf{z} \in \mathbb{C}^M$

マイクロフォンでの観測信号

Green関数を要素に持つ辞書行列

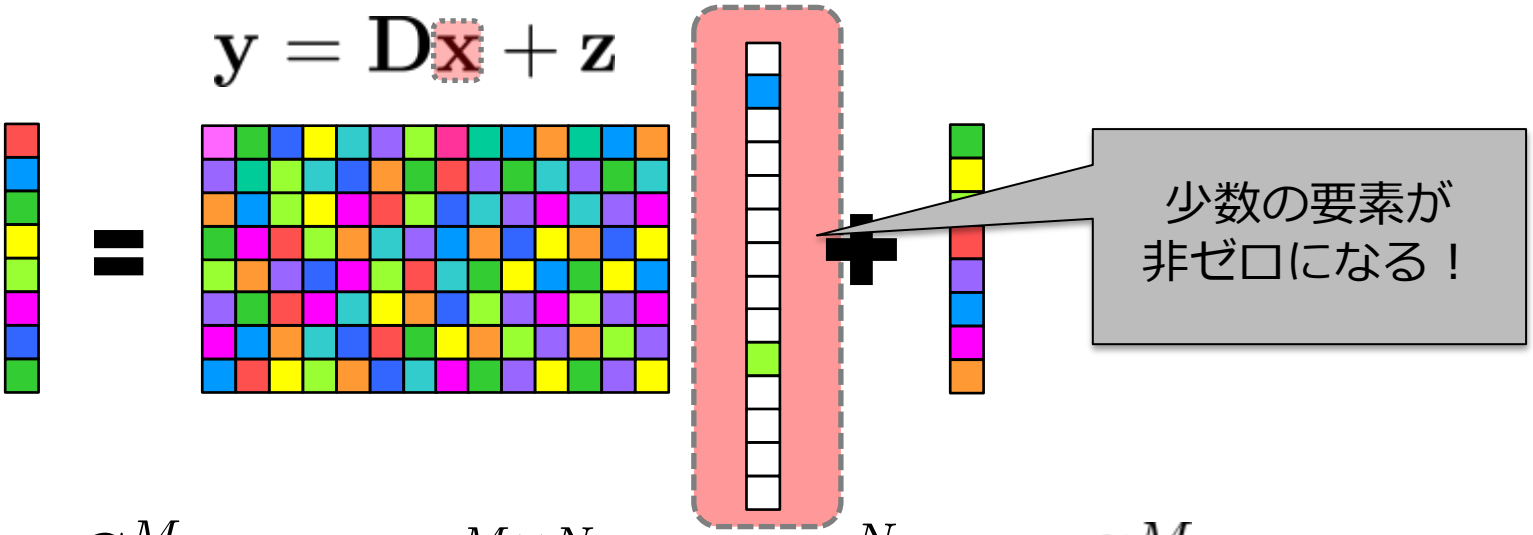
モノポール音源成分の分布

平面波由来の成分

スパース表現に基づく音場分解

モノポール音源成分が空間的にスパースであることを利用し，スパース分解の枠組みで求解

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{z}$$



少数の要素が非ゼロになる！

$\mathbf{y} \in \mathbb{C}^M$　$\mathbf{D} \in \mathbb{C}^{M \times N}$　$\mathbf{x} \in \mathbb{C}^N$　$\mathbf{z} \in \mathbb{C}^M$

最適化問題としてのスパース分解問題

$$\underset{\mathbf{x}}{\text{minimize}} \quad ||\mathbf{x}||_p$$

$$\text{subject to} \quad ||\mathbf{y} - \mathbf{D}\mathbf{x}||_2^2 < \varepsilon$$

$$0 < p \leq 1$$

$\mathbf{X}$ のスパース性を誘導するノルム

[2015小山]　Zの項にも低ランク制約

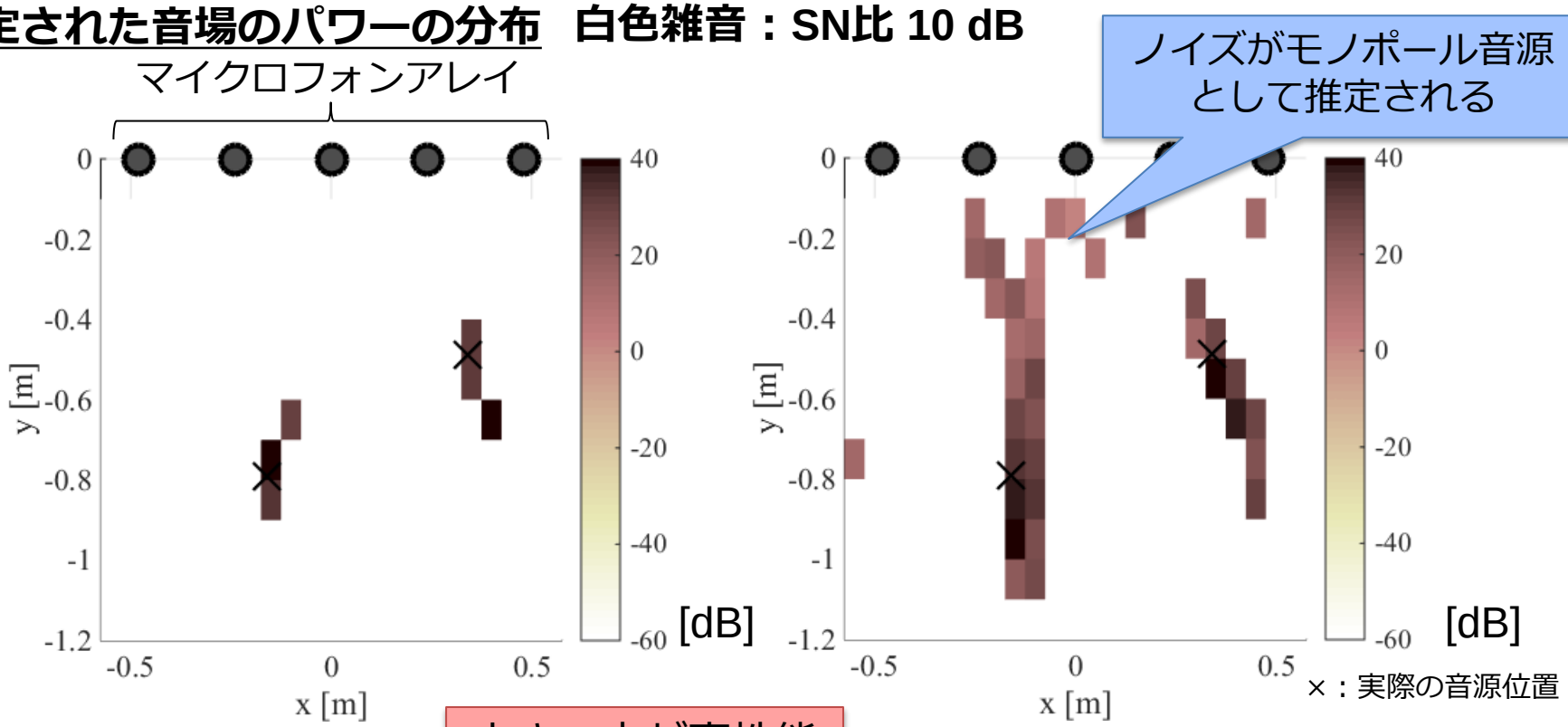
[2015村田]　音源自体を低ランク表現

ASJ学生優秀発表賞受賞！

シミュレーション実験：音源位置推定結果

強い音源の相関，また強いノイズの環境下において頑健な性能を示した.

推定された音場のパワーの分布　**白色雑音：SN比 10 dB**



提案法

大きい方が高性能

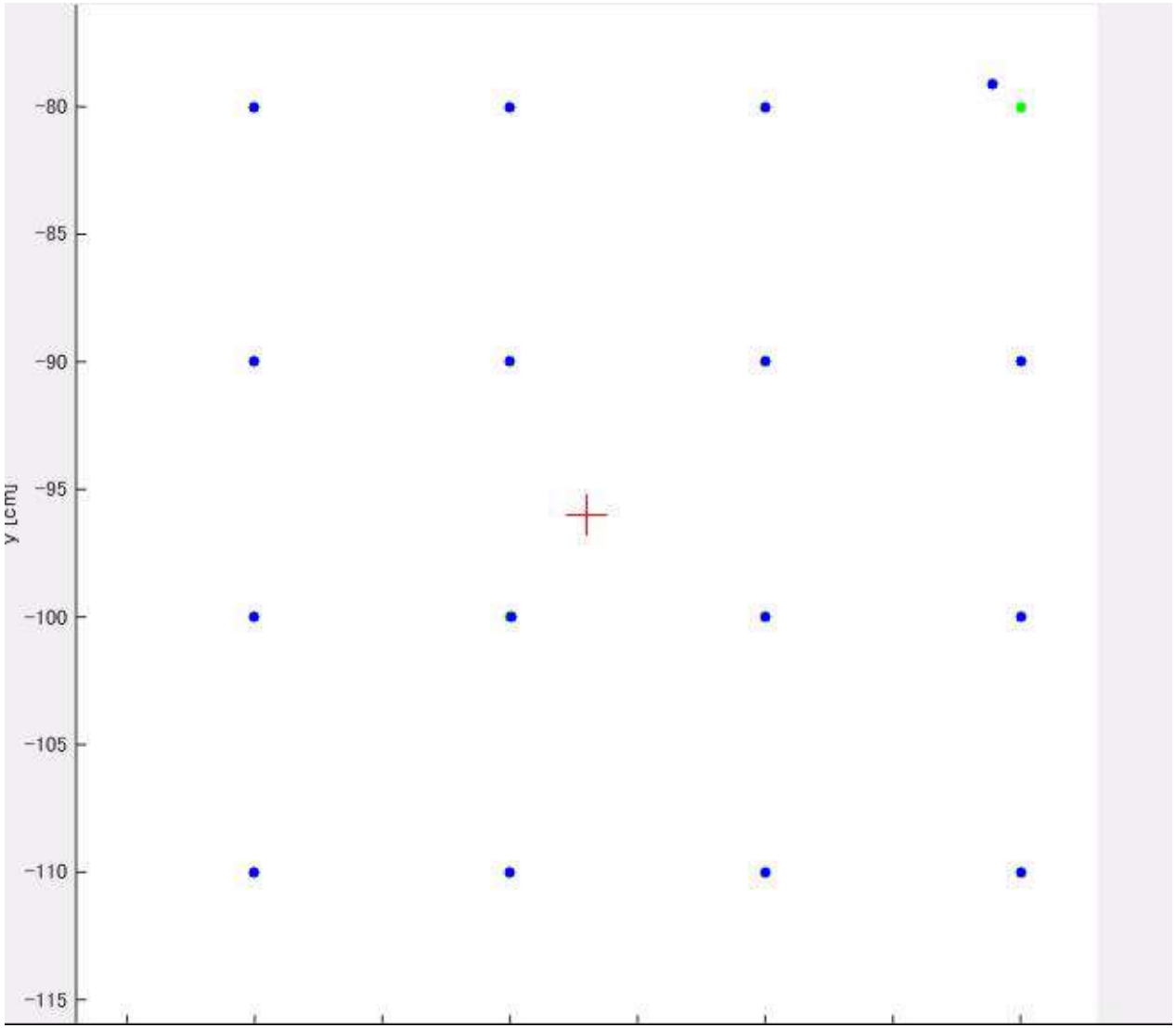
F-measure: 0.44

従来手法（M-FOCUSS）

F-measure: 0.25

音源位置とサンプリング点の同時推定結果

赤＋：真の音源，青●：サンプリング点



研究紹介５．統計的音声合成

- テキストもしくは自分の声を入れるだけで、誰の声でも、どんな訛りでも、何語でも、喋ることが出来るような統計的音声合成システムを実現する。

- 深層学習（Deep Neural Net）の枠組みを活かし、「AIオレオレ詐欺師」と「AI防犯課刑事」を対決させて、お互いに精度を高めるAnti-Spoofing 敵対学習理論を独自に提唱

ビッグデータ
深層学習
敵対学習

音声合成の統計モデリング

“おはよう” → テキスト音声合成 → 

x Probability$(y|x)$ y

入力 x と出力 y の関係をどう記述するか？ → 統計的逆問題

声の<u>ゆらぎ</u>をどう扱うか？

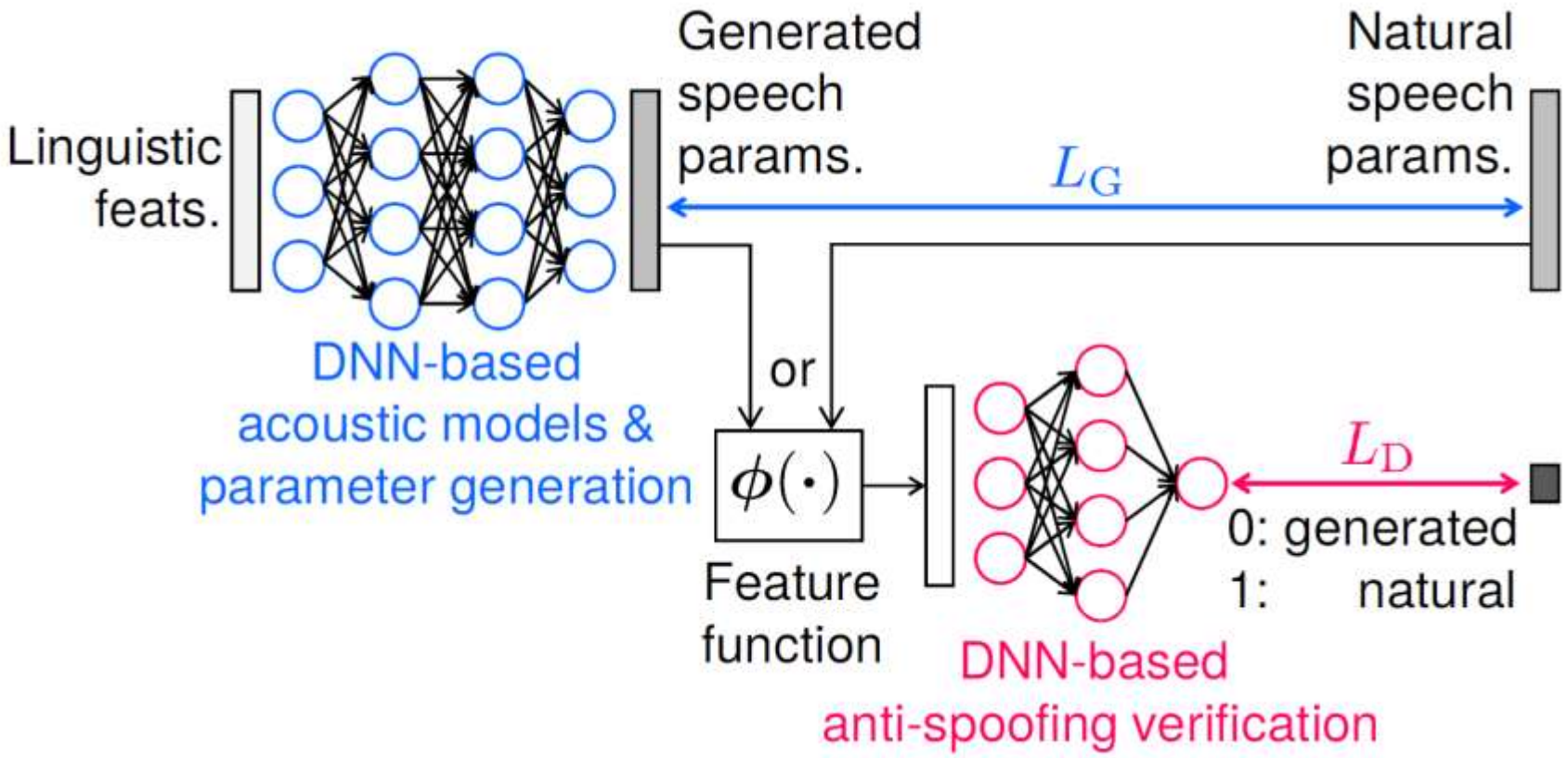
人間らしい声とは何か？「人間らし

2015年
テキスト音声合成国際コンペ一位
2016年
音声変換国際コンペ一位

東京大学は世界一！
昼飯はとんこつラーメンに限る！

Anti-Spoofingと敵対する音響モデル学習理論

[ICASSP2017 SLP Student Grant賞・IEEE SPSJ Student Paper Award]



人間の声に似せようと努力

ウソ（合成音）に騙されまいと攻防

統計制約と最適化

- **敵対的DNN音声合成 → anti-spoofing と音響モデルの交互最適化**

  Anti-spoofing の学習: cross-entropy 最小化

$$L_{\mathrm{D}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\frac{1}{T}\sum_{t=1}^{T}\log D(\boldsymbol{y}_t) - \frac{1}{T}\sum_{t=1}^{T}\log\left(1 - D(\boldsymbol{y}_t)\right)$$

  音響モデルの学習: MGE学習 + anti-spoofing との敵対的学習

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = L_{\mathrm{G}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \omega_{\mathrm{D}}\frac{E_{L_{\mathrm{G}}}}{E_{L_{\mathrm{D}}}} \times \left(-\frac{1}{T}\sum_{t=1}^{T}\log D(\hat{\boldsymbol{y}}_t)\right)$$

  敵対的学習: $\boldsymbol{y}$ と $\hat{\boldsymbol{y}}$ が従う確率分布間のJSダイバージェンス最小化
  [Goodfellow et al., 2014]

  コンテキストの条件付き敵対学習（音では世界初！）

- **異なる学習規範を用いた Generative Adversarial Network (GAN)**

  f-GAN [Nowozin et al., 2016]

    KL, JSダイバージェンスの拡張

  Wasserstein GAN (W-GAN) [Arjovsky et al., 2017]

    Wasserstein 距離 (earth mover's distance) の最小化

  分布に対して異なる感度の制約が加わる

- **積極的に自然音の分布を取り込むmoment-matching-DNN**

様々なGANによる声質の違い

識別器の違いに関する主観評価結果